Metadata Quality Evaluation: Experience from the Open Language Archives Community

Baden Hughes

Department of Computer Science and Software Engineering,
University of Melbourne,
Parkville VIC 3010, Australia
badenh@cs.mu.oz.au

Abstract. We describe the motivation, design and implementation of an infrastructure to support metadata quality assessment within a specialised Open Archives Initiative (OAI) sub-domain, the Open Language Archives Community (OLAC). While services for structural validation of metadata are widely used, there is little corresponding work regarding services which evaluate the semantic and syntactic content of metadata from a qualitative perspective. We posit that any measure of metadata quality benefits from both contextual and referential assessment - metadata on a per record and per collection basis is legitimately assessed against the baseline of broader community practice, as well as for compliance to any external standard. In this paper we describe the implementation of a metadata quality assessment scheme, and the corresponding interfaces to the evaluation tool.

1 Introduction

Much effort has been contributed to the design and implementation of metadata standards in the digital libraries community. The promotion of a distributed model of metadata creation and management, leveraged by central services for harvesting and aggregation of metadata, have resulted in a rapid expansion of the number of institutions and individuals who now contribute metadata to the digital libraries community as a whole. One consequence of the devolution of metadata creation and management has been that while metadata standards such as Dublin Core [1] are well accepted by the community as a whole, compliance with metadata standards and their domain-specific extensions is in fact highly variable.

While there are a number of structural validation tools available for metadata repositories within the Open Archives Initiative (OAI) [2] framework, relatively little effort has been expended on a framework for metadata quality assessment, a task which requires both structural and semantic validation and comparison.

In this paper, we report the motivation, design and implementation of an infrastructure to support metadata quality assessment within a specialised OAI sub-domain, the Open Language Archives Community (OLAC) [3]. We argue that the determination of metadata quality benefits from contextual and referential assessment: metadata on a per record and per collection basis is legitimately assessed against the baseline of broader community practice, as well as for compliance to external

Z. Chen et al. (Eds.): ICADL 2004, LNCS 3334, pp. 320–329, 2004.

standards. The combination of both these assessment models provides a significant incentive for archive managers to improve the quality of their metadata.

The structure of this paper is as follows: we contextualise the work within OLAC; describe our motivation in terms of the broader digital archives community; report the design of a metadata quality evaluation service; and review the implementation itself. Evaluation is followed by a brief discussion of items for future work. Finally we reflect on the broader implications of metadata quality evaluation for the digital libraries community as a whole.

2 Background

The Open Language Archives Community (OLAC) is a consortium of linguistic data archives, at the time of writing consisting of 29 archives and a corresponding catalogue of 27,000 objects described by metadata. (For a more detailed description of OLAC, we refer interested readers to [17] and [16]). OLAC metadata is based on Dublin Core, with a number of extensions [15] to the Dublin Core Metadata Set [4] for relevant conceptual domains such as language [10], linguistic type [7], subject language [8], linguistic subject [9] and linguistic role [11].

Derived from the model adopted within the OAI, the OLAC model has a two-tiered approach to implementation. Data providers are the institutional language archives which publish their XML-based metadata according to the OAI Static Repository standard [5]. Individual archives use a variety of software to manage their catalogues internally. Service providers leverage the OAI Protocol for Metadata Harvesting [6] to harvest the XML expressions of metadata catalogues. Within the OLAC community, typical practice is to aggregate these into an SQL database using the OLAC Harvester and Aggregator [12]. Service providers can then build services which utilise the union catalogue of OLAC metadata. (The work reported here is an example of exactly this type of service.)

As a metadata community and a virtual digital library, OLAC has motivated a number of developments at the OAI level, notably the need for supporting static repositories [5], the development of virtual service providers [13] and personal metadata creation and management tools [14].

3 Motivation

Our primary motivation in the work reported here is to establish infrastructural support at the service provider level to facilitate metadata quality assessment in an ongoing fashion. As mentioned earlier, one consequence of the devolution of metadata creation and management has been that while metadata standards such as Dublin Core are well accepted by the community as a whole, compliance with these metadata standards and their extensions is in fact highly variable.

While the structural validation tools for metadata (such as the OAI Repository Explorer [26]) are in use within OLAC, there is a notable absence of tools which are oriented towards both structural and semantic validation. Previous work within the OLAC context [27] has resulted in useful survey tools, although they lack a qualitative dimension. Spanne [29] provided a useful overview of the state of OLAC metadata after the first year of implementation experience, and in part motivates the

work reported here. In the broader OAI context, there are few examples of metadata quality assessment services (see for example [24], [25]). In particular, we draw motivation from previous work of Ward [19], where a longitudinal evaluation of the quality of DC metadata based on element and attribute usage within the OAI community is described.

Our work differs from previous efforts in three areas. First, we seek to establish a baseline against which future metadata instances can be compared in order to evaluate the maturity of the metadata creation process within OLAC. Second, we desire to provide assistance to individual data providers within OLAC with the means of self-evaluation and self-improvement. A third, but not insignificant goal, is to evaluate a number of domain-grounded controlled vocabularies specifically with regard to their adoption.

4 Algorithm Design

The objective of the algorithm is to arrive at a per metadata record score of between 0 and 10, based on the adherence to best practice guidelines for the use of Dublin Core metadata elements and codes ("core elements"), and the OLAC domain-specific controlled vocabularies ("codes"). Operations here are on a per metadata record basis.

The main algorithm derives two values - a Code Existence Score and an Element Absence Penalty, and then weighted to provide a Per Metadata Record Weighted Aggregate. Subsequent derivative metrics are then obtained using this aggregate as a baseline.

4.1 Code Existence Score

The principle for the derivation of the Code Existence Score is that for each element which has an associated extension from an OLAC controlled vocabulary, there is a corresponding increase in the quality of the metadata (ie. it becomes more fine-grained). A nominal score of 1 point is thus attributed to the metadata record. In turn, this is converted into a proportion of elements which use codes against the number of total elements in a record which have an associated controlled vocabulary. Hence the Code Existence Score is equal to the number of elements containing the code attributes divided by the number of elements of a type associated with a controlled vocabulary in the record, a value between 0 and 1.

4.2 Element Absence Penalty

The principle of deriving the Element Absence Penalty is that the quality of metadata declines with the absence of core elements which have been shown to be important to any metadata record (based on findings of Ward [19], Spanne [29] and our own surveys). Thus the following core elements have been deemed necessary in every record: title, description, subject, date and identifier.

For each of these core elements which is absent from a metadata record, a score of 0.2 is deducted from the metadata record score. This implies equal weighting of all the core elements. Hence, the Element Absence Penalty is equal to the number of core elements absent divided by the total number of core elements in a record, a value between 0 and 1.

4.3 Per Metadata Record Weighted Aggregate

Now we have both a Code Existence Score and an Element Absence Penalty, we combine these to derive a Per Metadata Record Weighted Aggregate. Essentially our approach is to reduce the theoretical maximum metadata record score of 10 by a factor proportional to the product of the Code Existence Score and the Element Absence Penalty. Hence the Per Metadata Record Weighted Aggregate is equal to the maximum score multiplied by the weighted product of the Code Existence Score and the Element Absence Penalty.

This results in an integer score out of 10 for each metadata record. These scores are held in a table ranking each item with a score out of 10. Scores are re-calculated when incremental metadata harvesting by the OLAC Harvester and Aggregator updates the collection of metadata records.

4.4 Derivative Metrics

Following this, a number of different metrics pertinent to metadata quality within an archive can be derived:

- Archive Diversity metric: A calculation of the diversity of controlled vocabulary usage within an archive, calculated for both subject and type. Diversity is calculated as being equal to the number of distinct code values divided by the number of instances of a metadata element, multiplied by 100 to provide a percentage.
- Metadata Quality Score metric: Derived from the aggregation of Per Metadata Record Weighted Aggregate scores within an archive.
- Core Elements Per Record metric: The percentage of records which have n of the core elements present at least once.
- Core Element Usage metrics: The percentage of records which contain the named element(s) at least once.
- Code Usage metrics: The number of times an element which has an associated code attribute is used by the archive, and the percentage of those elements which actually use a code attribute.
- Code and Element Usage metrics: The number of times an element is used. Where applicable, the number of times that a code attribute is used with that element.
- "Star Rating": A gross indicator, derived based on the average item score for the archive. It is calculated by dividing the average Per Metadata Record Weighted Aggregate for an archive by a factor of 2 and then applying rounding.

On the basis of this algorithm we can compute a score for each metadata record within an archive, for each archive in total, and for the community as a whole.

5 Implementation

The metadata quality assessment service is built on top of the foundational layer provided by the OLAC Harvester and Aggregator [12]. The implementation uses the open source technologies MySQL [22] and PHP [23], and is able to be installed on a range of platforms. The operational instance of the metadata quality assessment

service can be viewed online [21]. The software has been released under an open source license, and is freely available to interested parties from [20].

6 Evaluation

Using the metadata quality assessment infrastructure, we can now evaluate a number of different aspects of metadata quality within the OLAC context. We consider our findings in a number of areas: on a per data provider basis; across the whole community; trends and similarities between archives; and specifically within the OLAC context, the use of controlled vocabularies.

6.1 Evaluating Metadata Quality on a Per Data Provider Basis

It is immediately apparent that there is a high degree of variability between individual data providers within OLAC. Some archives have very high per metadata record scores, while others are very low. While we did not specifically set out to create an "archive ranking system", it is clear that some archives have significantly better quality metadata than others.

There appears to be no systematic correlation between the size of an archive and its corresponding average Per Metadata Record Weighted Aggregate. While the archive which at the time of writing has the largest number of metadata records also has the highest average Per Metadata Record Weighted Aggregate, and the smallest archives have the lowest average Per Metadata Record Weighted Aggregate, the mid-sized archives are apparently random in distribution in terms of average Per Metadata Record Weighted Aggregates.

There does however, appear to be a positive correlation between the size of an archive and the average number of elements per metadata record within the archive. This can be accounted for by the fact that larger archives typically export OLAC metadata as a derivative of a larger, much richer, metadata catalogue. OLAC metadata elements are optional and optionally repeatable, and it is clear that larger archives have a strong tendency to repeat elements such as subject and type, an approach inherited from the richer non-OLAC metadata natively used in such archives.

6.2 Evaluating Metadata Quality on a Community-Wide Basis

Across the entire community, metadata quality is averaged, and as such, broad measures of metadata quality are subject to some bias on the basis of the ratio between number of metadata records per archive and the corresponding Per Metadata Record Weighted Aggregate. If a large archive was to leave the community there would be a significant effect on the metadata quality metrics not simply owing to the number of records in any one archive, but their corresponding average metadata score.

We find evidence to reinforce the findings of Ward [19] in relation to the metadata elements which were most frequently used. Furthermore, we can see that in the context of OLAC fact element usage can also be classified into 4 distinct classes.

In the first class is a single element: subject, which is used twice as often as the members of the next class. The second class consists of five elements: title, description, date, identifier and creator, which are used around half as often as

subject. A third class contains format, type, contributor, publisher and isPartOf. Beyond this, a fourth class of infrequently used elements accounts for the remaining 33 elements from the Dublin Core Metadata Set. This last class interestingly includes language, which we find suprising given the linguistic focus of the digital archives within OLAC itself.

6.3 Qualitatively-Based Archive Clustering

Archives which have high metadata quality scores overall can be characterised as larger archives with high degree of quality control in the metadata creation process (generally long-standing archives with extensive infrastructural support, and for whom OLAC metadata is automatically generated). These archives have a tendency to use only core elements, and utilise sub-domain specific controlled vocabularies focused around subject. These archives have average Per Metadata Record Weighted Aggregate scores of between 8 and 10 points.

A second group of archives, characterised as smaller in size, but still having access to automated metadata generation systems archives, and significantly increased use of controlled vocabularies, can also be observed. These archives have average Per Metadata Record Weighted Aggregate scores between 4 and 7 points.

A final group of archives with the lowest average Per Metadata Record Weighted Aggregate scores (0-3 points) are clustered at the bottom of the list. These archives which provide only minimal metadata for small numbers of records, and their metadata catalogues have a tendency to be manually maintained. However, these archives are distinguished by virtue of their specialised holdings.

Over time, our objective is to promote upward migration on this scale - increasing the average item scores for each archive. One of the advantages of the infrastructure reported here is the fact that we can derive a longitudinal perspective on the evolution of metadata quality, a point we return to in the next section.

6.4 Use of OLAC Controlled Vocabularies

Specifically within the OLAC context, we can also evaluate the use of the various controlled vocabularies which distinguish this community from other Dublin Core-based efforts. Significant effort has been invested in the development of these controlled vocabularies, and the return on investment in these areas can now be assessed. In addition to raw statistics, we can observe a number of interesting trends across all archives.

In reference to subject, the controlled vocabulary was used 56% of the time, largely for language identification where international standards such as ISO-639 [28] (the recommended Dublin Core approach to language classification) are insufficiently granular to account for linguistic diversity. This contrasts with the use of the code for language, which is only used 30% of the time.

An interesting comparison can be made between creator and contributor. The same controlled vocabulary, the OLAC Role vocabulary [11] is applicable in both cases as an extension. However, we see that the creator code is used less that 1% of the time, while the contributor code is used 78% of the time. We can attribute this distinction perhaps to the process of creating language resources - there is typically one creator (a linguist) but multiple contributors (informants, translators etc). An additional factor here is that the Dublin Core recommendation to use contributor

rather than creator emerged just prior to 2 of the largest OLAC data providers joining the community.

Perhaps suprisingly, the type code is only used 33% of the time. The OLAC Linguistic Data Type vocabulary is only small (consisting of three elements), yet important from the perspective of linguists in distinguishing the various data types prevalent in language documentation and description. Despite the low volume of usage of this controlled vocabulary, we can adduce that much differentiation between linguistic data types can be performed based on the title and subject elements.

7 Future Work

Now we have established viable infrastructural support for metadata quality assessment within the OLAC domain, we now turn to a discussion of future work. We identify a number of natural extensions to the work reported here: an improvement to the algorithm; a longitudinal study of metadata quality; the need for new services which leverage the metadata quality assessment mechanism; possible new metrics; and an assessment methodology for human-agent collaboration in the metadata creation task.

One possible weakness in the algorithm used in the first instance is that no consideration is given to the size of the archive in determining its ranking. As an item of future work, we propose to extend the overall algorithm to derive a new metric, a weighted aggregate ranking, which will eliminate this possible weakness in the scheme as implemented currently.

A stable but dynamic infrastructure for metadata quality assessment allows us to conduct a longitudinal study of the changing nature of metadata implementation within OLAC. We propose to collate snapshots of the quality of metadata on a per data provider and the aggregate for the whole community on a monthly basis, and to archive these online as a precursor to trend analysis.

Based on the ranking system on a per record and per archive basis, we can conceive of integrating the metadata quality assessment data as a key part of a range of extended services. Perhaps the most immediately realisable of these is the use of metadata quality metrics to provide visual and logical ordering to search engine results. We have already commenced work on a general search engine for OLAC [18] which leverages domain specific knowledge, and we view the metadata quality assessment framework as an integral part of such a service.

A second generation of metadata quality evaluation metrics within the OLAC context is also emerging. While the instantiation reported here forms a useful starting point, other metrics which reflect the core values of OLAC (such as the availability of data online and the use of OLAC's fine-grained domain-specific vocabularies) are envisaged with subsequent alterations to the core algorithm. Additionally exploration of the consistency of application of metadata across a data provider's records could be considered, allowing insight at the level of individual records rather than through the Per Metadata Record Weighted Aggregate. Such a metric (which is oriented at the quality of an archive's metadata) could be contrasted with another which reflected the quality of an archive's collection.

Furthermore, we are conscious that in many cases, metadata creation is an human-intensive process, and as such any computational assistance which can be offered is

welcome. To date, one limiting factor has been the absence of an objective metadata quality evaluation framework. it is difficult to assess the contribution of pro-active metadata creation by computational agents. In the OLAC context, we can now explore the automated enrichment of existing metadata and the creation of new metadata based on web data mining approaches as discussed in [30], and assess the effectiveness of such approaches against baseline.

8 Conclusion

Our work here has resulted in the deployment of a scalable, dynamically-adjusted metadata quality evaluation infrastructure. In turn this allows a new perspective on the quality of metadata within digital archives. While the lack of infrastructural support for qualitative assessment of metadata within the digital archives community is notable, we believe that the provision of tools which assist metadata creators and managers to understand the qualitative aspects of metadata are of critical importance. Such tools enable archive managers to identify specific areas for metadata enrichment activity, and hence derive the greatest return on investment. Having now implemented of a framework for metadata quality assessment which is sustainable over an extended period, we hope that such a service will assist archive maintainers to focus their metadata improvement efforts, resulting in higher quality metadata across the whole Open Language Archives Community. We offer our approach and implementation to the broader digital libraries community in the hope that the model and implementation may benefit a larger range of institutional data providers, and ultimately, end-users.

Acknowledgements

We are grateful to Amol Kamat for his programming assistance; to Steven Bird for editorial comments on an earlier version of this paper; and to Gary Simons for informative discussions.

The work reported in this paper has been sponsored by the National Science Foundation Grant Numbers 9910603 (ISLE: International Standards in Language Engineering) and 0094934 (Querying Linguistic Databases).

References

1. Dublin Core. http://dublincore.org
2. Open Archives Initiative. http://www.openarchives.org
3. Open Language Archives Community. http://www.language-archives.org
4. Dublin Core Metadata Element Set, Version 1.1: Reference Description. http://dublincore.org/documents/dces
5. Patrick Hochstenbach, Henry Jerez, Herbert Van de Sompel, 2003. The OAI-PMH Static Repository and Static Repository Gateway. Proceedings of the IEEE/ACM Joint Conference on Digital Libraries 2003 (JDCL'03). pp. 210-220.

6. Carl Lagoze, Herbert Van de Sompel, Michael Nelson and Simeon Warner, 2002. The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html
7. Helen Aristar-Dry and Heidi Johnson, 2002. OLAC Linguistic Data Type Vocabulary. http://www.language-archives.org/REC/type.html
8. Gary Simons and Steven Bird, 2003. OLAC Subject Language Vocabulary. http://www.language-archives.org/REC/language.html
9. Helen Aristar-Dry and Michael Appleby, 2003. OLAC Linguistic Subject Vocabulary. http://www.language-archives.org/REC/field.html
10. Gary Simons and Steven Bird, 2003. OLAC Language Vocabulary. http://www.language-archives.org/REC/language.html
11. Heidi Johnson, 2003. OLAC Role Vocabulary. http://www.language-archives.org/REC/role.html
12. Gary Simons, 2003. A Query Facility for the Selective Harvesting of OLAC Metadata. http://www.language-archives.org/NOTE/query.html
13. Gary Simons, 2003. OLAC Virtual Service Provider. http://www.language-archives.org/viser
14. Kurt Maly, Mohammad Zubair and Xiaoming Liu, 2001. Kepler. An OAI Data/Service Provider for the Individual. D-Lib Magazine 7(4). http://www.dlib.org/dlib/april01/maly/04maly.html
15. Gary Simons and Steven Bird, 2002. Recommended Metadata Extensions. http://www.language-archives.org/REC/olac-extensions.html
16. Gary Simons and Steven Bird, 2003. The Open Language Archives Community: An infrastructure for distributed archiving of language resources. Literary and Linguistic Computing 18, pp.117-128.
17. Steven Bird and Gary Simons, 2003. Extending Dublin Core Metadata to support the description and discovery of language resources. Computing and the Humanities 37, pp.375-388.
18. Baden Hughes and Amol Kamat, 2004. A Metadata Search Engine for Language Archives. Manuscript.
19. Jewel Ward, 2003. A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative. Proceedings of the IEEE/ACM Joint Conference on Digital Libraries 2003 (JDCL'03). pp.315-317.
20. Open Language Archives Community Project at Sourceforge. http://olac.sourceforge.net
21. OLAC Archive Report Card. http://www.language-archives.org/tools/reports/archiveReportCard.php
22. MySQL Database Engine. http://www.mysql.com
23. PHP Scripting Engine. http://www.php.net
24. Lloyd Sokvitne, 2000. An Evaluation of the Effectiveness of Current Dublin Core Metadata for Retrieval. Proceedings of VALA 2000. Victorian Association for Library Automation: Melbourne.
25. Jane Greenberg, Maria Cristina Pattuelli, Bijan Parsia, and W. Davenport Robertson, 2001. Author-Generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organisation. Journal of Digital Information 2(2).
26. OAI Repository Explorer. http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai
27. OLAC Archive Survey. http://www.language-archives.org/tools/survey.php4

28. ISO, 1998. ISO 639-2: Codes for the representation of names of languages -- Part 2: Alpha-3 code. International Organisation for Standardization.
29. Joan Spanne, 2002. OLAC: The State of the Archives. Proceedings of the IRCS Workshop on Open Language Archives. Institute for Research in Cognitive Science, University of Pennsylvania. pp.42-46.
30. Baden Hughes, 2004. Perspectives on Metadata. Proceedings of the LREC 2004 Workshop on Building the Language Resources and Evaluation Roadmap: Joint COCOSDA and ICCWLRE Meeting. European Language Resources Association: Paris.